Présentation générale de l'Analyse en Composantes Principales (ACP)

Cette présentation reprend des éléments du polycopié sur l'ACP, qui utilise lui même des éléments d'un polycopié rédigé par Marie-Christine Roubaud (I2M). Elle reprend en partie et complète le premier cours sur l'ACP normée.

Nous ne ferons pas, ni en cours ni en TP, la plupart des preuves des résultats. Néanmoins les étudiants sont encouragés à travailler les démonstrations en travail autonome, en se limitant par exemple

aux deux ACP typiques : canonique et normée,

et au cas d'individus équi-pondérés.

La plupart des démonstrations reposent sur des calculs algébriques accessibles.

Partie 1 : Définitions

On observe p variables quantitatives X^1, X^2, $\ldots$, X^p sur n individus.

Matrice X ayant n lignes (**individus**) et p colonnes (**variables**) :

$$X = \begin{pmatrix} x_1^1 & x_1^2 & \ldots & x_1^p \\ x_2^1 & x_2^2 & \ldots & x_2^p \\ \vdots & \vdots & \vdots & \vdots \\ x_n^1 & x_n^2 & \ldots & x_n^p \end{pmatrix}$$

où x_i^j : valeur prise par la j ème variable sur le $i$ème individu.

On associe à chaque individu i un point x_i de $\mathbb{R}^p$, à savoir transposée de la ième ligne de la matrice X :

$$x_i = \begin{pmatrix} x_i^1 \\ x_i^2 \\ \vdots \\ x_i^p \end{pmatrix}$$

On associe à chaque variable X^j le vecteur x^j de $\mathbb{R}^n$, à savoir la $j$ème colonne de la matrice X :

$$x^j = \begin{pmatrix} x_1^j \\ x_2^j \\ \vdots \\ x_n^j \end{pmatrix}$$

Abus de langage par la suite : x_i=individu i, x^j=variable j.

On affecte à chaque individu i un **poids** p_i : $p_i > 0$ et $\sum_{i=1}^{n} p_i = 1$.

Matrice des poids :

$$D = \begin{pmatrix} p_1 & 0 & \dots & 0 \\ 0 & p_2 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & p_n \end{pmatrix}$$

Nuage des individus : points x_i munis de leurs poids :

$$\mathcal{M} = \{(x_i, p_i)\,;\ i = 1, \dots, n\}\,.$$

Rappels :

- Moyenne (empirique) de la variable x^j : $\quad \bar{x}^j = \sum_{i=1}^n p_i x_i^j$.
- Variance (empirique) de x^j :
  $s_j^2 = \sum_{i=1}^n p_i (x_i^j - \bar{x}^j)^2 = \sum_{i=1}^n p_i (x_i^j)^2 - (\bar{x}^j)^2$.
- Covariance (empirique) de x^j et x^k :
  $s_{jk} = \sum_{i=1}^n p_i (x_i^j - \bar{x}^j)(x_i^k - \bar{x}^k) = \sum_{i=1}^n p_i x_i^j x_i^k - \bar{x}^j \bar{x}^k$.
- Corrélation (empirique) de x^j et x^k : $\quad r_{jk} = \frac{s_{jk}}{s_j s_k}$.

Soit V la matrice de covariance et R la matrice de corrélation, qui joueront un rôle important par la suite :

$$V = \begin{pmatrix} s_1^2 & s_{12} & \dots & s_{1p} \\ s_{21} & s_2^2 & \dots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \dots & s_p^2 \end{pmatrix},$$

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix}.$$

Centre de gravité ou (barycentre) du nuage de points $\mathcal{M}$:

$$g = {}^t(\bar{x}^1, \bar{x}^2, \cdots, \bar{x}^p)$$

Variable centrée :

$$y^j = {}^t(x_1^j - \bar{x}^j, x_2^j - \bar{x}^j, \ldots, x_n^j - \bar{x}^j) = x^j - \bar{x}^j {}^t(1, \ldots, 1)$$

Individu centré y_i : $y_i = {}^t(x_i^1 - \bar{x}^1, x_i^2 - \bar{x}^2, \cdots, x_i^p - \bar{x}^p) = x_i - g$

Tableau des données centrées $y_i^j = x_i^j - \bar{x}^j$:

$$Y = \left[y^1, y^2, \cdots, y^p\right] = \begin{pmatrix} y_1^1 & \cdots & y_1^j & \cdots & y_1^p \\ y_2^1 & \cdots & y_2^j & \cdots & y_2^p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_n^1 & \cdots & y_n^j & \cdots & y_n^p \end{pmatrix}$$

Nuage centré des individus :

$$\mathcal{N} = \{(y_i, p_i)\,;\ i = 1, \ldots, n\}$$

Variables centrées et réduites $z^1, \ldots, z^p$:

$$\forall i \in \{1, \ldots, n\},\ z_i^j = \frac{x_i^j - \bar{x}^j}{s_j}.$$

Tableau des données centrées et réduites :

$$Z = \left[z^1, z^2, \cdots, z^p\right] = \begin{pmatrix} z_1^1 & \ldots & z_1^j & \ldots & z_1^p \\ z_2^1 & \ldots & z_2^j & \ldots & z_2^p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ z_n^1 & \ldots & z_n^j & \ldots & z_n^p \end{pmatrix}.$$

Individus centrés et réduits $z_1, \ldots, z_n$: transposées des lignes de Z

Remarque : propriétés sur colonnes de Y ou Z (centrées et réduites),
mais aucune propriété sur les lignes de Y ou Z !

Matrice diagonale des inverses des écarts-types :

$$D_{1/S} = diag(1/s_1, \cdots, 1/s_p).$$

Matrice diagonale des inverses des variances :

$$D_{1/S^2} = diag(1/s_1^2, \cdots, 1/s_p^2) = D_{1/S} \times D_{1/S}.$$

Proposition

1) $Z = YD_{1/S}$

2) $V = {}^t XDX - g\,{}^t g = {}^t YDY$

3) $R = D_{1/S} V D_{1/S} = {}^t ZDZ$

Comment définir une distance entre deux individus ?

Métrique générale sur $\mathbb{R}^p$:

$$\forall a, b \in \mathbb{R}^p,\ d_M(a,b) = \sqrt{{}^t(a-b)M(a-b)} = ||a-b||_M\,,$$

où M est une matrice symétrique de taille (p,p) définie positive

Produit scalaire sur $\mathbb{R}^p$ associé :

$$< a, b >_M = {}^t aMb.$$

Cas particulier avec "orthogonalité" des axes :

$$d(a,b) = \sqrt{m_1(a^1-b^1)^2 + m_2(a^2-b^2)^2 + \cdots + m_p(a^p-b^p)^2}\,,$$

où $m_1 > 0,\ldots,m_p > 0$

Deux cas particuliers courants :

1) Travailler sur les données Y avec $M = I_p$, **ACP canonique ou simple**.
   Utilisation : variables ont la même unité de mesure et de même ordre de grandeur.

2) Travailler sur les données Y avec $M = D_{1/S^2}$, **ACP normée (1ère version)**
   Utilisation : variables n'ont pas la même unité de mesure ou ne sont pas de même ordre de grandeur.

ACP normée souvent plus simple à interpréter que ACP canonique.

On peut présenter ACP normée d'une autre façon (quasiment) équivalente :

2') Travailler sur les données Z avec $M = I_p$, **ACP normée (2ème version)**

1ère version : permet de présenter ACP normée comme cas particulier d'une théorie générale, utilisée dans la suite de la présentation...
...Mais en pratique on pourra préférer la 2ème version, pour sa simplicité.
(petite différence entre les 2 versions pour objets u^k et d_k définis plus tard)

Comment calculer une distance dans l'espace des variables $\mathbb{R}^n$?

Le choix de la métrique $D = diag(p_1, p_2, \ldots, p_n)$ est intéressant. En effet :

- $< y^j, y^k >_D = s_{jk}$
- $||y^j||_D^2 = s_j^2$
- $\cos\theta_{jk} = \frac{<y^j, y^k>_D}{||y^j||_D ||y^k||_D} = \frac{s_{jk}}{s_j s_k} = r_{jk}$, avec θ_{jk} angle entre les vecteurs y^j et y^k

Et on a des propriétés analogues pour toutes les variables centrées de $\mathbb{R}^n$.

Les résultats qui suivent sont écrits dans un cas général pour M et D, mais dans une 1ère approche vous pouvez vous limiter à

$$M = I_p$$

$$D = \frac{1}{n} I_n$$

Et à des données qui sont soit Y (ACP canonique) soit Z (ACP normée),
et la matrice "VM" à diagonaliser plus loin est soit V (ACP canonique) soit R (ACP normée)

Définition : Inertie totale du nuage des individus

$$I = \sum_{i=1}^{n} p_i d_M^2(x_i, g) = \sum_{i=1}^{n} p_i ||x_i - g||_M^2 .$$

Exemple :

Soit $\mathcal{M}$ le nuage des cinq points suivants de $\mathbb{R}^2$:

$$A(1,0) \quad B(-1,0) \quad C(0,0), \quad D(-1,1), \quad E(0,2) ,$$

à qui on attribue le même poids.

1) On considère la distance euclidienne usuelle sur $\mathcal{M}$. Faire un graphique représentant ces points dans un repère orthonormé et calculer l'inertie du nuage.

2) *Changement de métrique :* déterminer l'expression analytique de la distance d_M sur $\mathbb{R}^2$ associé à la matrice $M = D_{1/S^2}$. Calculer l'inertie dans ce cas.

L'inertie du nuage $\mathcal{M}$ est égale à l'inertie du nuage centré $\mathcal{N}$

$$I = \sum_{i=1}^{n} p_i ||x_i - g||_M^2 = \sum_{i=1}^{n} p_i ||y_i||_M^2 = \sum_{i=1}^{n} p_i d_M^2(y_i, O)\,.$$

Ainsi 1ère phase de l'ACP = centrage des variables

Proposition : $I = tr(VM)$.

Conséquence :

- ACP canonique : $I = \sum_{i=1}^{p} s_j^2$
- ACP normée $I = p$

Inertie du nuage des individus $\mathcal{N}$ expliquée par un sous-espace vectoriel F de $\mathbb{R}^p$:

$$I_F = \sum_{i=1}^{n} p_i ||\hat{y}_i^F||_M^2 ,$$

où $\hat{y}_i^F$ = projection orthogonale de y_i sur F.

Cas particulier $F = \Delta_u$, la droite vectorielle engendrée par u un vecteur M-normé, i.e $||u||_M = 1$:
Alors $\hat{y}_i^u$ la projection orthogonale de y_i sur Δ_u est

$$\hat{y}_i^u = < y_i , u >_M \ u,$$

et l'inertie expliquée par Δ_u est

$$I_{\Delta_u} = \sum_{i=1}^{n} p_i ||\hat{y}_i^u||_M^2 = {}^t u M V M u.$$

Propriété : Si $F = F_1 \oplus F_2$ et $F_1 \perp F_2$ alors

$$I_F = I_{F_1} + I_{F_2}.$$

Conséquence : Pour un sev F donné on a $\mathbb{R}^p = F \oplus F^\perp$, donc

$$I = I_F + I_{F^\perp}.$$

$I_{F^\perp}$ peut-être considérée comme la déformation du nuage projeté sur F :

$$I_{F^\perp} = \sum_{i=1}^{n} p_i ||y_i - \hat{y}_i^F||_M^2.$$

Partie 2 : Formulation du problème de l'ACP et solution

Objectif principal : obtenir une représentation "fidèle" du nuage des individus de $\mathbb{R}^p$ en le projetant sur un espace de faible dimension k_0.

Choix de l'espace de projection fait selon le critère de l'inertie :

- choisir une valeur pour k_0 (voir partie 5)
- chercher un (le ?) sous-espace de dimension k_0 expliquant le maximum d'inertie possible
- de façon équivalente : chercher un (le ?) sous-espace de dimension k_0 qui engendre la déformation la plus faible possible

En général, à k_0 fixé, la solution est unique. Le sev solution, noté E_k, est nommé espace principal de dimension k_0.

Théorème :

$$E_{k+1} = E_k \oplus \Delta_{u_{k+1}}$$

où $\Delta_{u_{k+1}}$ droite vectorielle portant l'inertie maximale parmi toute les droites M-orthogonales à E_k,
(et où on a noté u_{k+1} un vecteur M-orthonormé qui engendre cette droite.)

D'où procédure pour trouver E_k :

- Rechercher un axe Δ_{u_1} maximisant l'inertie expliquée $I_{\Delta_{u_1}} = {}^t u_1 MVMu_1 \rightarrow$ on obtient $E_1 = \Delta_{u_1}$.
- Rechercher un axe Δ_{u_2} orthogonal à E_1, maximisant l'inertie expliquée $I_{\Delta_{u_2}} = {}^t u_2 MVMu_2 \rightarrow E_2 = E_1 \oplus \Delta_{u_2}$.
- ...
- Rechercher un axe Δ_{u_k} orthogonal à E_{k-1} maximisant l'inertie expliquée $I_{\Delta_{u_k}} = {}^t u_k MVMu_k \rightarrow E_k = E_{k-1} \oplus \Delta_{u_k}$.

Définition : matrice $A(p,p)$ M-symétrique

$$\forall a, b \in \mathbb{R}^p, \ < a \, , \, Ab >_M = < Aa \, , \, b >_M$$

Conséquence du théorème spectral : Toute matrice $A(p,p)$ réelle M-symétrique admet p valeurs propres réelles et une base de vecteurs propres M-orthonormés.

Hypothèse pour toute la suite : Y est de rang p (vrai en général dans les applications)

Théorème
Soit une base M-orthonormée $(e_1, e_2, \cdots, e_p)$ formée par les vecteurs propres de la matrice VM rangés par ordre décroissant des valeurs propres $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Alors $\lambda_p > 0$ et les vecteurs $u_1 = e_1$, $u_2 = e_2$, $\ldots$, $u_k = e_k$ engendrent les axes recherchés pour construire E_k.

Définition:

- $u_1, \ldots, u_p$ vecteurs principaux
- $\Delta_{u_1}, \ldots, \Delta_{u_p}$ axes principaux
- rappel : $E_k = \Delta_{u_1} \oplus \cdots \oplus \Delta_{u_k}$ espace principal de dimension k

Propriété : $I_{\Delta_{u_j}} = \lambda_j$

On en déduit $I_{E_k} = \lambda_1 + \cdots + \lambda_k$

Remarque : En prenant $k = p$, on retrouve
$I = \sum_{j=1}^{p} \lambda_j = tr(VM)$.

à retenir :

- ACP *canonique ou simple* : analyser données Y, métrique des individus $M = I_p$ (identité)
  $\rightarrow$ ACP repose sur diagonalisation de V

- ACP *normée* : deux façons de faire (quasi) identiques

  soit analyser données Y, métrique des individus $M = D_{1/S^2}$
  $\rightarrow$ ACP repose sur diagonalisation de VM

  soit analyser données Z, métrique des individus $M = I_p$
  $\rightarrow$ ACP repose sur diagonalisation de R

Partie 3 : Analyse du nuage des individus

Décomposition du vecteur y_i sur la base des vecteurs principaux $(u_1, u_2, \cdots, u_p)$:

$$y_i = \sum_{j=1}^{p} c_i^j u_j$$

où $c_i^j =< y_i, u_j >_M = {}^t y_i M u_j$

Définition:
jème composante principale : $c^j = {}^t(c_1^j, c_2^j, \ldots, c_n^j)$
càd coordonnées sur u_j du nuage $\mathcal{N}$ projeté orthogonalement sur Δ_{u_j}

Proposition : $c^j = YMu_j$

Conséquences :

- chaque CP est une combinaison linéaire des variables de départ $y^1, \ldots, y^p$,
- chaque CP c^j est centrée, et de variance λ_j,
- Les CP sont non corrélées deux à deux.

Remarque : Ainsi la matrice de variance-covariance des nouvelles variables c^j est diagonale.

Carte des individus :

Pour tout $k \leq p$ et $l \leq p$, $k \neq l$,
on a (c_i^k, c_i^l) = coordonnées sur u_j et u_k du projeté orthogonal de y_i sur le plan $\Delta_{u_k} \oplus \Delta_{u_l}$

Et $\{(c_i^k, c_i^l)\,;\, i = 1, \cdots, n\}$ coordonnées sur u_k et u_l du nuage $\mathcal{N}$ projeté orthogonalement sur $\Delta_{u_k} \oplus \Delta_{u_l}$

Représentation graphique sur le plan, avec c_i^k en abscisse et c_i^l en ordonnée, est dite "carte des individus sur le plan principal (k, l)"

Remarque : Pour deux individus h et i on a
$d_M(y_h, y_i) = \|y_h - y_i\|_M = \sqrt{\sum_{j=1}^{p}(c_h^j - c_i^j)^2}$

- si y_h et y_i sont dans le plan principal (k, l) :

$$d_M(y_h, y_i) = \sqrt{(c_h^k - c_i^k)^2 + (c_h^l - c_i^l)^2}$$

utiliser la distance canonique du plan sur la carte des individus donne $d_M(y_h, y_i)$.

- Si y_h et y_i ne sont pas dans le plan principal (k, l) on a :

$$d_M(y_h, y_i) \geq \sqrt{(c_h^k - c_i^k)^2 + (c_h^l - c_i^l)^2}.$$

utiliser la distance canonique du plan sur la carte des individus sous-estime les distances $d_M(y_h, y_i)$,
carte intéressante seulement pour individus "bien représentés" (voir plus loin).

Rappel inertie totale du nuage $\mathcal{N}$: $I = \sum_{j=1}^{p} \lambda_j = tr(VM)$

Définition:

- Qualité globale de la repésentation du nuage $\mathcal{N}$ sur le s.e principal E_k :

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{I} = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\sum_{j=1}^{p} \lambda_j}.$$

- Qualité de représentation de l'individu i sur l'axe principal Δ_{u_k} :

$$cos^2(y_i, \hat{y}_i^{u_k}) = \frac{||\hat{y}_i^{u_k}||_M^2}{||y_i||_M^2} = \frac{(c_i^k)^2}{\sum_{j=1}^{p} (c_i^j)^2}$$

cosinus carré de l'angle que fait y_i avec $\hat{y}_i^{u_k}$ sa projection sur Δ_{u_k}

Interprétation :

- Si $\cos^2(y_i, \hat{y}_i^{u_k})$ proche de 1, individu i bien représenté sur Δ_{u_k}
- Si $\cos^2(y_i, \hat{y}_i^{u_k})$ est proche de 0, individu i mal représenté sur Δ_{u_k}

Généralisation possible à un sous-espace E_k, exemple qualité de représentation de l'individu i sur le premier plan principal E_2 :

$$\cos^2(y_i, \hat{y}_i^{E_2}) = \frac{||\hat{y}_i^{E_2}||_M^2}{||y_i||_M^2} = \frac{(c_i^1)^2 + (c_i^2)^2}{\sum_{j=1}^{p}(c_i^j)^2} = \cos^2(y_i, \hat{y}_i^{u_1}) + \cos^2(y_i, \hat{y}_i^{u_2})$$

On a par définition :

$$I = \sum_{i=1}^{n} p_i ||y_i||_M^2.$$

Définition:

- $p_i \|y_i\|_M^2 = p_i \sum_{k=1}^{p} (c_i^k)^2$ **contribution** de l'individu i à l'inertie totale,
- $\frac{p_i \|y_i\|_M^2}{I} = \frac{p_i \sum_{k=1}^{p} (c_i^k)^2}{I}$ **contribution relative** de l'individu i à l'inertie totale.

On a vu que $I_{\Delta_{u_k}} = \lambda_k$, or on a

$$\lambda_k = Var(c^k) = \sum_{i=1}^{n} p_i (c_i^k)^2$$

Définition:

- $p_i (c_i^k)^2$ **contribution** de l'individu i à l'inertie portée par Δ_{u_k}
- $\frac{p_i (c_i^k)^2}{\lambda_k}$ **contribution relative** de l'individu i à l'inertie portée par Δ_{u_k}

Partie 4 : Analyse du nuage des variables

Nuage des variables :

$$\mathcal{V} = \{y^1, y^2, \ldots, y^p\}.$$

Rappel : on utilise métrique des poids D pour calculer des distances entre éléments de $\mathcal{V}$, ou plus généralement entre des vecteurs centrés de $\mathbb{R}^n$.

Remarque : on n'introduit pas de "barycentre" des variables. Point important est l'origine de $\mathbb{R}^n$.

Rappel Inertie totale :

$$I = tr({}^tYDYM).$$

On voit que c'est aussi :

$$I = tr(YM{}^tYD),$$

semblable 1ère expression en inversant les rôles : les "individus" sont les variables, la "métrique" est D et la "pondération" est M.

Ainsi on peut formuler un problème "d'ACP sur les variables" analogue au problème d'ACP sur les individus : trouver les sous-espaces F^k de $\mathbb{R}^n$ qui conservent au mieux l'inertie.

Nom consacré : "factoriel" (rappel pour les individus : "principal")
On ne détaillera pas la formulation et résolution du problème "d'ACP sur les variables", on donne les définitions des objets "factoriels", qui sont liés aux objets "principaux".

Rappel : on garde l'hypothèse $rang(Y) = p$

Proposition Diagonalisation de YM^tYD :

- p valeurs propres positives non nulles $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ et $n - p$ valeurs propres nulles (avec les mêmes λ_k que ceux introduits dans la section "individus")
- Les CP $(c^1, c^2, \ldots, c^p)$ sont des vecteurs propres resp. associés aux valeurs propres $\lambda_1, \ldots, \lambda_p$.

Rappel : CP sont $D-$orthogonales, mais pas $D-$normées ($var(c^j) = \lambda_j$). On en déduit les "facteurs" qui sont une base $D-$orthonormée de $\mathcal{V}$.

Définition:

- $\forall\ 0 \leq k \leq p$, $v^k = \frac{1}{\sqrt{\lambda_k}} c^k$ kième facteur
- $\forall\ 0 \leq k \leq p$, $\Delta_{v^k} = Vect(v^k)$ kième axe factoriel
- $F^k = Vect\{v^1, v^2, \cdots, v^k\}$ espace factoriel de dimension k

Remarque : v^k est en fait c^k réduite, mais en général v^k pas interprétée comme une variable

Comme dans le domaine des individus, on peut définir une inertie expliquée I' par un sev de l'espace des variables $\mathbb{R}^n$. On admet alors :

Proposition

$I'_{\Delta_{v^k}} = \lambda_k$

$I'_{F^k} = \lambda_1 + \cdots + \lambda_k$

On définit $d_1, d_2, \ldots, d_p$ les "Composantes Principales de l'ACP sur les variables" :
d_k=vecteur de $\mathbb{R}^p$ contenant coordonnées des variables $y^1, y^2, \ldots, y^p$ sur v^k

$$d_k^j = < y^j, v^k >_D \quad \text{càd} \quad d_k = {}^t Y D v^k$$

Proposition : $d_k = \sqrt{\lambda_k} u_k$, pour tout $k = 1, \ldots p$

Décomposition du vecteur y^j sur la famille des facteurs $(v^1, v^2, \cdots, v^p)$:

$$y^j = \sum_{k=1}^{p} d_k^j v^k .$$

Definition

Pour tout $k \leq p$ et $l \leq p$, $k \neq l$, la représentation des points $\{(d_k^j, d_l^j)\,;\, j = 1, \cdots, p\}$ dans le plan, avec d_k^j en abscisse et d_l^j en ordonnée, est appelé **carte des variables sur le plan factoriel** (k, l).

Remarque 1

utiliser des vecteurs partant de l'origine pour représenter les variables : meilleure visualisation des normes, produits scalaires et angles, qui jouent un rôle important

Remarque 2 :

On a $< y^j, y^{j'} >_D = \sum_{k=1}^{n} d_k^j d_k^{j'}$

Ainsi si y^j et $y^{j'}$ sont dans le plan factoriel (k, l) :

$$< y^j, y^{j'} >_D = d_k^j d_k^{j'} + d_l^j d_l^{j'} = < (d_k^j, d_l^j), (d_k^{j'}, d_l^{j'}) >,$$

avec $<, >$ produit scalaire canonique de $\mathbb{R}^2$

Ainsi produits scalaires (et donc aussi les normes et angles) des variables initiales
= produits scalaires (et normes et angles) sur la carte (k, l) avec la géométrie classique de $\mathbb{R}^2$

Par contre faux pour variables mal représentées, notamment normes vues sur la carte < vraies normes des variables initiales.

Définition:

- $\frac{\lambda_1+\cdots+\lambda_k}{\sum_{j=1}^{p}\lambda_j}$ qualité globale de la représentation du nuage $\mathcal{V}$ sur F^k

- $\cos^2(y^j,\hat{y}^{j,v^k}) = \frac{||\hat{y}^{j,v^k}||_D^2}{||y^j||_D^2} = \frac{<y^j,v^k>_D^2}{s_j^2} = r^2(y^j,c^k)$ qualité de représentation de y^j sur le kième axe factoriel

On en déduit qualité de représentation de y^j sur tout plan factoriel.
Exemple sur le premier plan factoriel :

$$\cos^2(y^j, \hat{y}^{j,F_2}) = \frac{||\hat{y}^{j,F_2}||_D^2}{||y^j||_D^2} = r^2(y^j, c^1) + r^2(y^j, c^2)$$

Interprétation :

- Si $\cos^2(y^j, \hat{y}^{j,F_2}) \simeq 1$: y^j bien représentée dans F_2
- Si $\cos^2(y^j, \hat{y}^{j,F_2}) \simeq 0$: y^j mal représentée sur F_2

Remarques sur la notion de contribution d'une variable à une inertie

On a vu que $d_k = \sqrt{\lambda_k} u_k$, et donc $||d_k||^2_M = \lambda_k$.

Toutefois, pour M générale, relation ne donne pas une somme de contributions par variable.

Sauf si $M = Diag(m_1, \ldots, m_p)$ (ex ACP canonique ou normée), relation donne :

$$\lambda_k = \sum_{j=1}^{p} m_j (d_k^j)^2$$

Ainsi $m_j (d_k^j)^2$ sorte de contribution de la variable j à l'inertie portée par l'axe factoriel k,

et $\frac{m_j (d_k^j)^2}{\lambda_k}$ sorte de contribution relative.

Spécificités de l'ACP normée

Prenons 2ème version : on fait l'ACP simple des données $z^j = \frac{x^j - \bar{x}^j}{s_j}$

calcul des CP de l'ACP sur les variables. On remarque que

$$d_k^j =< z^j, v^k >_D = r(z^j, v^k)$$

or $v^k = \frac{1}{\sqrt{\lambda_k}} c^k$, donc $d_k^j = r(z^j, v^k) = r(z^j, c^k)$

Donc d_k^j = corrélation entre la variable initiale z^j et la nouvelle variable synthétique c^k

En fait on voit que quantités $d_k^j = r(z^j, c^k)$ jouent un rôle majeur dans l'analyse des variables en ACP normée :

d_k^j coordonnée de z^j sur le kème axe factoriel,

d_k^j corrélation entre z^j et c^k,

$(d_k^j)^2$ qualité de représentation z^j sur le kème axe factoriel,

$(d_k^j)^2$ (respectivement $\frac{(d_k^j)^2}{\lambda_k}$) sorte de contribution (resp contribution relative) de la variable j à l'inertie portée par le kème axe factoriel.

Carte des variables : cercles des corrélations

Soit $\hat{z}^j$ projection $D-$orthogonale de z^j sur le plan factoriel (k, l).

On a $d_D(0, z^j) = ||z^j||_D = 1$

Conséquence 1 :

$z^1, \ldots, z^p$ appartiennent à hypersphère de $\mathbb{R}^n$ de centre 0 et de rayon 1,
donc $||\hat{z}^j||_D^2 = (d_k^j)^2 + (d_l^j)^2 \leq 1$: carte des variables à l'intérieur du cercle de $\mathbb{R}^2$ de centre 0 et de rayon 1, dit "cercle des corrélations"

Conséquence 2 :

Qualité : $\cos^2(\hat{z}^j, z^j) = ||\hat{z}^j||_D^2 = (d_k^j)^2 + (d_l^j)^2$
Donc variable j bien représentée sur le plan factoriel (k, l) si sa représentation sur la carte est proche du cercle des corrélations.

Rappel : géométrie classique de $\mathbb{R}^2$ utile sur une carte des variables

Pour deux variables initiales z^j et $z^{j'}$ (si bien représentées) :

- proximité sur la carte = forte corrélation linaire entre z^j et $z^{j'}$
- diamétralement opposés sur la carte = corrélation négative proche de -1,
- des directions presque orthogonales sur la carte = faible corrélation

Entre variable initiale z^j et nouvelles variables c^k et c^l

- Si sur la carte, $\hat{z}^j$ proche du point $(1,0)$ (respectivement $(-1,0)$) : z^j est très corrélée positivement (resp négativement) avec c^k.
- Si sur la carte, $\hat{z}^j$ proche du point $(0,1)$ (respectivement $(0,-1)$) : z^j est très corrélée positivement (resp négativement) avec c^l.
- Si sur la carte, $\hat{z}^j$ proche de l'origine, alors z^j est peu corrélée avec c^k et c^l.

On a aussi liens entre la carte des variables sur (k, l) et la carte des individus sur (k, l).

Individus situés le plus à droite sur la carte des individus ont en général :

- valeurs fortement positives pour les variables initiales à droite de la carte des variables
- valeurs fortement négatives pour les variables initiales à gauche de la carte des variables

Inversement pour les individus les plus à gauches sur la carte des individus.

Idem individus situés en haut ou en bas de la carte des individus liés avec les variables initiales situées en haut ou en bas de la carte des variables.

Partie 5 : Pratique de l'ACP

Nombre d'axes à retenir

De nombreux critères de choix pour k_0 existent.

Exemple de la règle de Kaiser :

Rappel : on a $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$,
$Inertie = \lambda_1 + \lambda_2 + \cdots + \lambda_p > 0$

Choisir le plus grand k_0 tel que $\lambda_{k_0} \geq \frac{Inertie}{p}$

Définition:
On appelle **élément supplémentaire** : élément (individu ou variable) qui n'a pas été pris en compte dans l'analyse pour la détermination des axes, mais dont on a calculé ensuite ses coordonnées sur chacun des axes pour le porter sur les graphiques.

Intéressant pour interpréter les ACP :

- proximité d'un individu supplémentaire bien connu avec un groupe de points initiaux anonymes,

- proximité d'une variable supplémentaire avec une CP : corrélation instructive, qui vient uniquement de liens statistiques et non pas de "biais de construction"

Démarche possible pour décrire une carte des variables ou des individus :

1- Donner le pourcentage d'inertie expliquée par le plan considéré et chacun des axes,

2- Indiquer les variables (resp.les individus) mal représentées dans ce plan pour les exclure de la description,

3- Utiliser les *contributions*

- *des variables* (si bien définies) pour interpréter les axes en termes de variables de départ. Pour chaque axe factoriel important, lister variables initiales de plus forte contribution à cet axe, et les répartir dans un tableau (positif ou négatif) selon signe de leur coordonnée sur l'axe.
- *des individus* pour identifier ceux qui sont influents pour l'orientation d'un axe (en placer si besoin en individus supplémentaires)

4-a **Pour une carte des variables** : étudier les angles entre les projections des variables en termes de covariance ou de corrélation (selon le type d'ACP choisi) pour dégager éventuellement des *groupes* de variables.

4-b **Pour une carte d'individus** : étudier les proximités ou les oppositions entre les points en termes de "comportement" et dégager éventuellement des *groupes* d'individus

5- Faire une synthèse des informations et hypothèses principales dégagées de la carte décrite.

Effet de taille

Supposons variables initiales toutes corrélées positivement entre elles, cad matrice de variance-covariance et la matrice de corrélation ont tous leurs termes positifs.

Théorème d'algèbre (de Frobenius) : si ACP simple ou normée, alors 1ère composante principale corrélée positivement avec toutes les variables initiales,
donc les individus sont rangés sur le premier axe principal par valeurs croissantes de l'ensemble des variables (en moyenne).

On dit que c^1 définit un facteur de "taille",
et c^2 différencie alors les individus de "taille" similaires, facteur de "forme"